

Consensus Message Passing for Layered Graphical Models

Varun Jampani[†]
MPI for Intelligent Systems, Tübingen

S. M. Ali Eslami[†], Daniel Tarlow, Pushmeet Kohli and John Winn
Microsoft Research, Cambridge

Abstract

Generative models provide a powerful framework for probabilistic reasoning. However, in many domains their use has been hampered by the practical difficulties of inference. This is particularly the case in computer vision, where models of the imaging process tend to be large, loopy and layered. For this reason bottom-up conditional models have traditionally dominated in such domains. We find that widely-used, general-purpose message passing inference algorithms such as Expectation Propagation (EP) and Variational Message Passing (VMP) fail on the simplest of vision models. With these models in mind, we introduce a modification to message passing that learns to exploit their layered structure by passing *consensus* messages that guide inference towards good solutions. Experiments on a variety of problems show that the proposed technique leads to significantly more accurate inference results, not only when compared to standard EP and VMP, but also when compared to competitive bottom-up conditional models.

1 Introduction

Generative models provide a powerful framework for probabilistic reasoning and are applicable across a wide variety of domains, including computational biology, natural language processing, and computer vision. For example, in computer vision, one can use graphical models to express the process by which a face is lit and rendered into an image, incorporating knowledge of surface normals, lighting and even the approximate symmetry of human faces. Models that make effective use of this information will generalize well, and they will require less labelled training data

[†]The first two authors contribute equally to this work.

than their unstructured counterparts (*e.g.* random forests or neural networks) in order to make accurate predictions.

Perhaps the most significant challenge of the generative modelling framework is that inference can be very hard. Sampling-based methods run the risk of slow mixing, while message passing-based methods (which are the focus of this work) can converge slowly, converge to bad solutions, or fail to converge at all. Whilst significant efforts have been made to improve the accuracy of message passing algorithms (*e.g.* by using structured variational approximations), many challenges remain, including difficulty of implementation, the problem of computational cost and the question of how the structured approximation should be chosen. The present work aims to alleviate these problems for general-purpose message-passing algorithms.

Our starting observation was that general purpose message passing inference algorithms (*e.g.* EP and VMP; Minka, 2001; Winn and Bishop, 2005) fail on even the simplest of computer vision models. We claim that in these models the failure can be attributed to the algorithms' inability to determine the values of a relatively small number of influential variables which we call 'global' variables. Without accurate estimation of these global variables, it can be very difficult for message passing to make meaningful progress on the other variables in the model.

Latent variables in vision models are often organised in a layered structure, where the observed image pixels are at the bottom and high-level scene parameters are at the top. Additionally, knowledge about the values of the variables at level l is sufficient to reason about any global variable at layer $l + 1$. With these properties in mind, we develop a method called *Consensus Message Passing* (CMP) that learns to exploit such layered structures and estimate global variables during the early stages of inference.

Experimental results on a variety of problems show that CMP leads to significantly more accurate inference results whilst preserving the computational efficiency of standard message passing. The implication of this work is twofold. First, it adds a useful tool to the toolbox of techniques for improving general-purpose inference, and second, in doing so it overcomes a bottleneck that has restricted the use of model-based machine learning in computer vision.

2 Consensus Message Passing

Consensus message passing exploits the layered characteristic of vision models in order to overcome the aforementioned inference challenges. For illustration, two layers of latent variables of such a model are shown in Fig. 1a using factor graph notation (black). Here the latent variables below ($\mathbf{h}^b = \{h_k^b\}$) are a function of the latent variables above ($\mathbf{h}^a = \{h_k^a\}$) and the global variables x and y (where k ranges over pixels; in this case $|k| = 3$). As we will see in the experiments that follow, this is a recurring pattern that appears in many models of interest in vision. For example, in the case of face modeling, the \mathbf{h}^a variables correspond to the normals \mathbf{n}_i , the global variable x to the light vector \mathbf{l} , and \mathbf{h}^b to the shading intensities s_i (see Fig. 6b).

Our reasoning follows a recursive structure. Assume for a moment that in Fig. 1a, the messages from the layer below to the inter-layer factors (blue) are both informative and accurate (e.g. due to being close to the observed pixels). We will refer to these messages collectively as *contextual messages*. It would be desirable, for purposes of both speed and accuracy, that we could ensure that the messages sent to the layer above (\mathbf{h}^a) also possess the same properties. If we had access to an oracle that could give us the correct belief for the global variables (x and y) for the image, we could send accurate initial messages from x and y and in one step compute informative and accurate messages from the inter-layer factors to the layer above.

In practice, however, we do not have access to such an oracle. In this work we train regressors to *predict* the values of the global variables given all the messages from the layer below. Should this prediction be good enough, the messages to the layer above will be informative and accurate, and the inductive argument will hold. We describe how these regressors are trained in Sec. 3. To summarize, the approach consists of the following two components:

1. Before inference, for each global variable in different layers of the model, we train a regressor to predict some oracle’s value for the target variable given the values of all the messages from the layer below (i.e. the *contextual messages*, Fig. 1a, blue),
2. During inference, each regressor sends this belief in the form of a *consensus message* (Fig. 1a, red) to its target variable.

In some models it will be useful to employ a second type of CMP, displayed graphically in Fig. 1b, where global layer variables are absent and loops in the graphical model are due to global variables in other layers. Here, a consensus message is sent to each variable in the latent layer above, given all the contextual messages.

Any message passing schedule can be used subject to the constraint that the consensus messages are given maximum

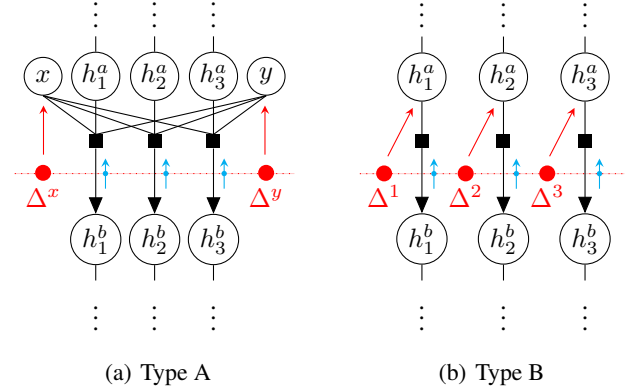


Figure 1: **Consensus message passing.** Vision models tend to be large, layered and loopy. (a) Two adjacent layers of the latent variables of a model of this kind (black). In CMP, consensus messages (red) are computed from contextual messages (blue) and sent to global variables (x and y), guiding inference in the layer. (b) Consensus message passing of a different kind for situations where loops in the graphical model are due to global variables in other layers.

priority within a layer and that they are sent bottom up. Naturally, a consensus message can only be sent once its contextual messages have been computed. It is desirable to be able to ensure that the fixed point reached under this scheme is also a fixed point of standard message passing in the model. One approach for this is to reduce the certainty of the consensus messages over the course of inference, or to only pass them in the first few iterations. In our experiments we found that even passing consensus messages only in the first iteration led to accurate inference, and therefore we follow this strategy for the remainder of the paper. It is worth emphasizing that message-passing equations remain unchanged and we used the same scheduling scheme in all our experiments (i.e. no need for manual tuning).

It is important to highlight a crucial difference between consensus message passing and heuristic *initialization*. In the latter, predictions are made from the *observations* no matter how high up in the hierarchy the target variable is, whereas in CMP predictions are made using *messages* that are sent from variables immediately below the target variables of interest. The CMP prediction task will be much simpler, since the relationship between the target variables and the variables in the layer immediately below is much less complex than the relationship between the target variables and the observations. Furthermore, we know from the layered structure of the model that all relevant information from the observations is contained in the variables in the layer below. This is because target variables at layer $l + 1$ are conditionally independent of all layers $l - 1$ and below, given the values of layer l .

One final note on the capacity of the regressors. Of course it is true that an infinite capacity regressor can make perfect predictions given enough data (whether using CMP or heuristic initialization). However we are interested in practical ways of obtaining accurate results for models of increasing complexity, where lack of capable regressors and unlimited data is inevitable. One important feature of CMP is that it makes use of predictors in a scalable way, since regressions are only made between adjacent latent layers.

3 Predictor Training

To recap, we wish to perform inference in a layered model of observed variables \mathbf{X} with latent variables \mathbf{H} . Each predictor Δ^t (with target t) is a function of a collection of its contextual messages $\mathbf{c} = \{c_k\}$ (incoming from the latent layer below \mathbf{h}^b), that produces the consensus message m , *i.e.* $m = \Delta^t(\mathbf{c})$.

We adopt an approach in which we *learn* a function for this task that is parameterized by θ , *i.e.* $\bar{m} \equiv f(\mathbf{c}|\theta)$. This can be seen as an instance of the canonical regression task. For a given family of regressors f , the goal of training is to find parameters θ that capture the relationship between context and consensus message pairs $\{(\mathbf{c}_d, m_d)\}_{d=1\dots D}$ in some set of training examples.

3.1 Choice of predictor training data

First we discuss how this training data is obtained. There can be at least three different sources:

1. Beliefs at convergence. This technique is only useful if standard message passing works but is slow. Standard message passing inference is run in the model for a large number of iterations and for a collection of different observations $\{\mathbf{X}_d\}$. Message passing is scheduled in precisely the same way as it would be if CMP were present, however no consensus messages are sent. For each observation \mathbf{X}_d , the collection of the marginals of the latent variables in the layer below the predictor ($\mathbf{h}_d^b = \{h_{dk}^b\}$, see *e.g.* Fig. 1a) at the *first* iteration of message passing is considered to be the context \mathbf{c}_d , and the marginal of the target variable t at the *last* iteration of message passing is considered to be the oracle message m_d . The aim is that during inference on new problems, a predictor trained in this way would send messages that *accelerate* convergence to the fixed-point that message passing would have reached by itself anyway.

2. Samples from the model. This technique is useful if standard message passing fails to reach good fixed points no matter how long it is run for. First a collection of samples from the model is generated, giving us for each sample both the observation \mathbf{X}_d and its corresponding latent variables \mathbf{H}_d . Standard message passing inference is then run on the observations $\{\mathbf{X}_d\}$ only for a single iteration. Message passing is scheduled as before. For each observation

\mathbf{X}_d , the marginals of the latent variables in the layer below \mathbf{h}_d^b at the *first* iteration of message passing is the context \mathbf{c}_d , and the oracle message m_d is considered to be a point-mass centered at the sampled value of the target variable t . The aim is that during inference on new problems, a predictor trained in this way would send messages that guide inference to a fixed-point in which the marginal of the target variable t is close to its sampled value.

3. Labelled data. As above, except the latent variables of interest \mathbf{h}_d are set from real data instead of being sampled from the model. The oracle message m_d is therefore a point-mass centered at the label provided for the target variable t for observation \mathbf{X}_d . The aim is that during inference on new problems, a predictor trained in this way would send messages that guide inference to a fixed-point in which the marginal of the target variable t is close to its labelled value, even in the presence of a degree of model mismatch. We demonstrate each of the strategies in the experiments in Sec. 4.

3.2 Random regression forests

We wish to learn a mapping f from contextual messages \mathbf{c} to the consensus message m from training data $\{(\mathbf{c}_d, m_d)\}_{d=1\dots D}$. This is challenging since the inputs and outputs of the regression problem are messages (*i.e.* distributions), and special care needs to be taken to account for this fact. We follow closely the methodology of Eslami et al. (2014), in which random forests are used to predict outgoing EP messages from a factor. A detailed description of our random forest implementation is provided in the supplementary material. For a review of forests see Criminisi and Shotton (2013).

4 Experiments

We first illustrate the application of CMP to two diagnostic models: one of circles and a second of squares. We then use the approach to improve inference in a more challenging vision model: that of intrinsic images of faces. In the first experiment the predictors are trained on beliefs at convergence, in the second on samples from the model, and in the third on annotated labels, showcasing various use-cases of CMP. We show that in all cases the proposed technique leads to significantly more accurate inference results whilst preserving the computational efficiency of message passing. The experiments were performed in Infer.NET (Minka et al., 2012) using default settings, unless stated otherwise. We set the number of trees in each forest to 8.

4.1 A generative model of circles

We begin by studying the behaviour of standard message passing on a simplified Gauss and Ceres problem (Teets and Whitehead, 1999). We use this example to highlight

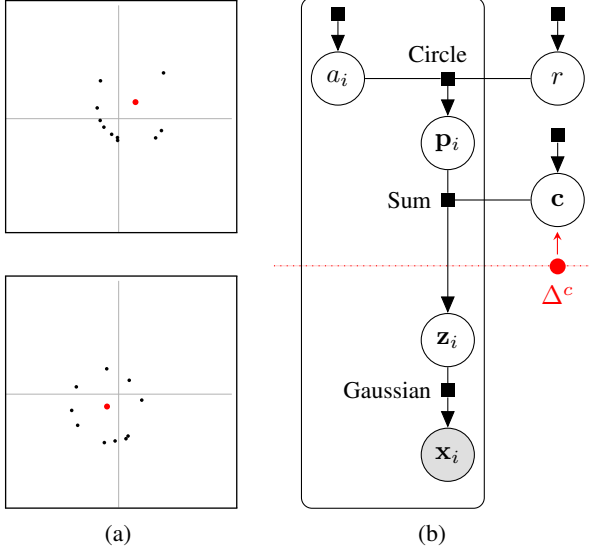


Figure 2: **The circle problem.** (a) Given a sample of points on a circle (black), we wish to infer the circle’s center (red) and its radius. Two sets of samples are shown. (b) The graphical model for this problem.

the fact that although inference may require many iterations of message passing, message initialization can have a significant effect on the speed of convergence, and to demonstrate how this can be done automatically using CMP.

Given a noisy sample of points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1\dots N}$ on a circle in the 2D plane (Fig. 2a, black, $\mathcal{N}(0, 0.01)$ noise on each axis), we wish to infer the coordinates of the circle’s center \mathbf{c} (Fig. 2a, red) and its radius r . We can express the data generation process using a graphical model (Fig. 2b). The Cartesian point $(0, r)$ is rotated a_i radians to generate \mathbf{p}_i , then translated by \mathbf{c} to generate the latent \mathbf{z}_i , which finally produces the noisy observation \mathbf{x}_i . This model can be expressed in a few lines of code in Infer.NET. The circle model is interesting for our purposes since it is both layered (the \mathbf{z}_i s, \mathbf{p}_i s and a_i s each form a layer) and loopy (due to the presence of two variables outside the plate).

Vanilla message passing inference in this model can take a surprisingly large number of iterations to converge. We draw 10 points $\{\mathbf{x}_i\}$ from circles with random centers and radii, run VMP and record the accuracy of the marginals of the latent variables at each iteration. We repeat the experiment 50 times and plot results in Fig. 3 (dashed black). As can be seen from the figure, the marginals contain significant errors even after 50 iterations of message passing.

We then experiment with consensus message passing. A predictor Δ^c is trained to send a consensus message to \mathbf{c} in the initial stages of inference, given the messages coming up from all of the \mathbf{z}_i (indicated graphically in Fig. 2b, red). The predictor is trained on final beliefs at 100 iterations of standard message passing on $D = 500$ sample problems.

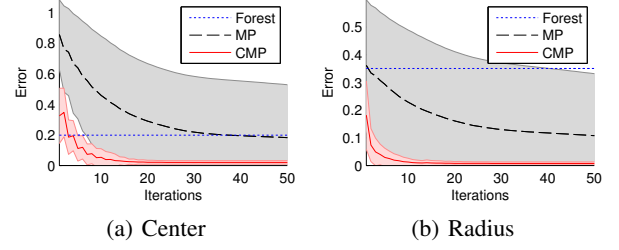


Figure 3: **Accelerated inference using CMP.** (a) Distance of the mean of the marginal posterior of c from its true value as a function of number of inference iterations (Forest: direct prediction, MP: standard VMP, CMP: VMP with consensus). Consensus message passing significantly accelerates convergence. (b) Similar plot for radius r .

As can be seen in Fig. 3 (red), this single consensus message has the effect of significantly increasing the rate of convergence (as indicated by slope) and also inference robustness (as indicated by error bars). For comparison we also plot how well a regressor of the same capacity as the one used by CMP can directly estimate the latent variables without using the graphical model in Fig. 3 (blue). Consensus message passing gives us the best of both worlds in this example: speed that is more comparable to one-shot bottom-up prediction and the accuracy of message passing inference in a good model for the problem.

4.2 A generative model of squares

Next we turn our attention to a more challenging problem for which even the best message passing scheme that we could devise frequently finds completely inaccurate solutions. The task is to infer the center \mathbf{c} and side length r of a square in an image (Fig. 4a). Unlike the previous problem where we knew that all points belonged to the circle, here we must first determine which pixels belong to the square and which do not. To do so we might also wish to reason about the colour of the foreground \mathbf{fg} and background \mathbf{bg} , making the task of inference significantly harder. The graphical model for this problem is shown in Fig. 4b.

We experiment with 50 test images (themselves samples from the model), perform inference using EP and with a sequential schedule, recording the accuracy of the marginals of the latent variables at each iteration. We additionally place damping with step size 0.95 on messages from the square factor to the center \mathbf{c} . We found these choices led to the best performing standard message passing algorithm. Despite this, we observed inference accuracy to be disappointingly poor (see Fig. 5). In Fig. 5a we see that, for many images, message passing converges to highly inaccurate marginals for the center. The low quality of inference can also be seen in quantitative results of Figs. 5(b-d).

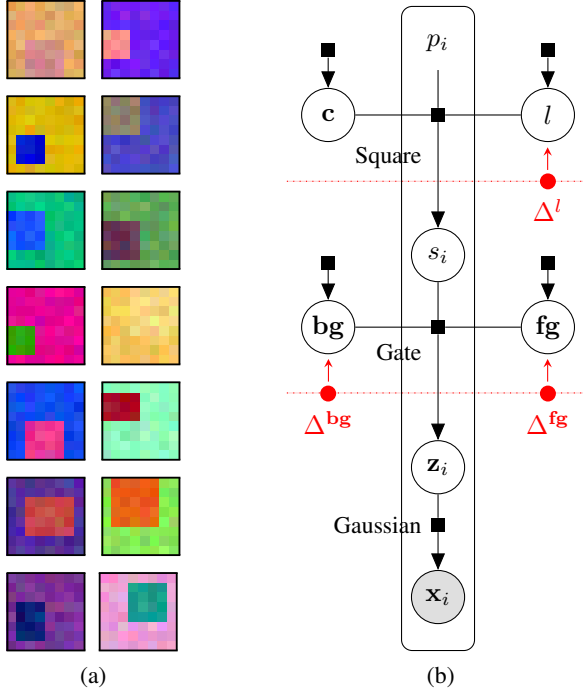


Figure 4: **The square problem.** (a) We wish to infer the square’s center and its side length. (b) A graphical model for this problem. s_i is a boolean variable indicating the square’s presence at position p_i . Depending on the value of s_i , the gate copies the appropriate colour (\mathbf{fg} or \mathbf{bg}) to \mathbf{z}_i .

We implement CMP predictors at two different layers of the model (see Fig. 4b, red). In the first layer, $\Delta^{\mathbf{fg}}$ and $\Delta^{\mathbf{bg}}$ send consensus messages to \mathbf{fg} and \mathbf{bg} respectively, given the messages coming up from all of the \mathbf{z}_i which take the form of independent Gaussians centered at the appearances of the observed pixels (we use a Gaussian noise model). Therefore $\Delta^{\mathbf{fg}}$ and $\Delta^{\mathbf{bg}}$ effectively make initial guesses of the values of the foreground and background colours in the image given the observed image. Split features in the internal nodes of the regression forest are designed to test for equality of two randomly chosen pixel positions, and sparse regressors are used at the leaves to prevent overfitting.

In the second layer, Δ^l sends a consensus message to l given the messages coming up from all of the s_i . The messages from s_i take the form of independent Bernoullis indicating the algorithm’s current beliefs about the presence of the square at each pixel. Therefore the predictor’s job is to predict the square’s side length from this probabilistic segmentation map. Note that it is much easier to implement a regressor to perform this task (effectively one only needs to count) than it is to do so using the original observed image pixels x_i . We find these predictors to be sufficient for stable inference and so we do not implement a fourth predictor for \mathbf{c} . We experiment with single stage CMP, where only the lower predictors $\Delta^{\mathbf{fg}}$ and $\Delta^{\mathbf{bg}}$ are active, and with

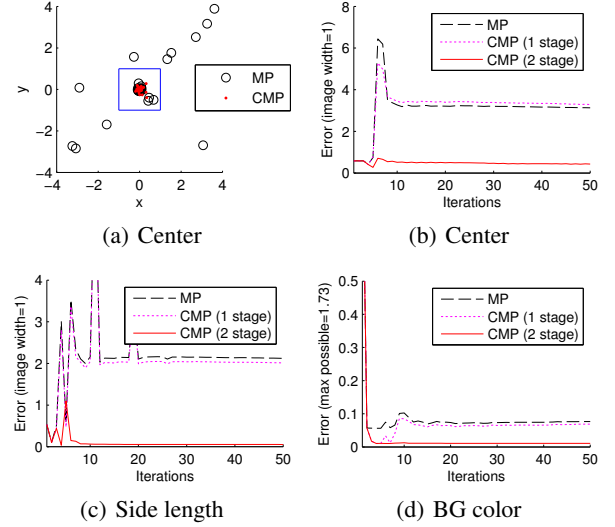


Figure 5: **Robustified inference using CMP.** (a) Position of inferred centers relative to groundtruth. Image boundaries shown in blue for scale. (b,c,d) Distance of the mean of the posterior of \mathbf{c} , l and \mathbf{bg} from their true values. CMP consistently increases inference accuracy. Results have been averaged over 50 different problems. 1 stage CMP only makes use of the lower predictors $\Delta^{\mathbf{fg}}$ and $\Delta^{\mathbf{bg}}$.

two stage CMP, where all three predictors are active. The predictors are trained on $D = 500$ samples from the model.

The results of these experiments are shown in Fig. 5. We observe that CMP significantly improves the accuracy of inference for the center \mathbf{c} (Figs. 5a, 5b) but also for the other latent variables (Figs. 5c, 5d). Of note is the fact that single stage CMP appears to be insufficient for guiding message passing to good solutions. Whereas in circle example CMP accelerated convergence, this example demonstrates how it can make inference possible in models that were outside the capabilities of standard message passing.

4.3 A generative model of faces

We also investigate a more realistic application to face modelling. The estimation of reflectance and shape from a single image of a human face is a well-studied problem in computer vision (see *e.g.* Georgiades et al. 2001; Lee et al. 2005; Wang et al. 2009; Kemelmacher-Shlizerman and Basri 2011; Tang et al. 2012). A primary motivation for this task is that reflectance and shape are invariant to confounding light effects, and are therefore useful for downstream tasks such as recognition. The problem is ill-posed however, and modern approaches make heavy use of prior knowledge in order to obtain good solutions, *e.g.* in the form of average reflectance and normal statistics (Biswas et al., 2009; Biswas and Chellappa, 2010) or morphable 3D models (Zhang and Samaras, 2006; Wang et al., 2009).

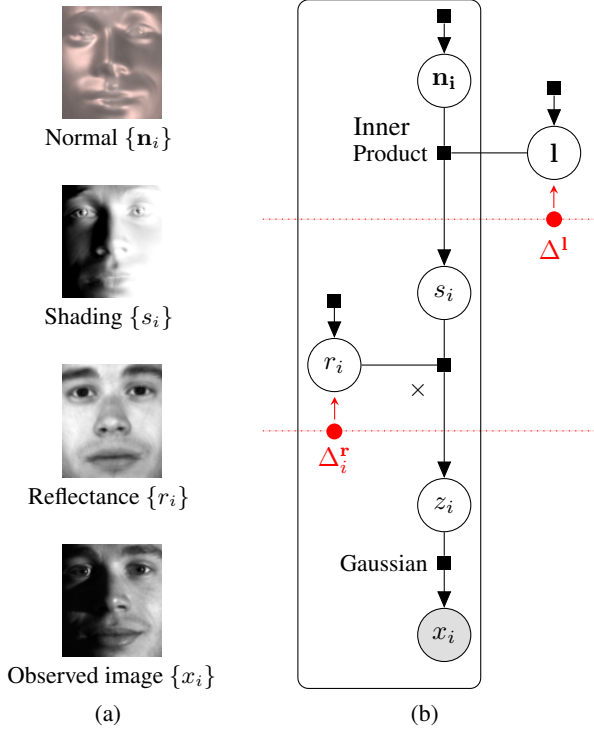


Figure 6: **The face problem.** (a) We observe an image and wish to infer the corresponding reflectance map and normal map (visualized here as 3D shape). (b) A graphical model for this problem. Symmetry priors not shown.

Model. Given an observation of pixels $\mathbf{x} = \{x_i\}$, we wish to infer the reflectance value r_i and normal vector \mathbf{n}_i for each pixel i (see Fig. 6a). In Fig. 6b, a model is shown for these variables that represents the following image formation process: $x_i = (\mathbf{n}_i \cdot \mathbf{l}) \times r_i + \epsilon$, thereby assuming Lambertian reflection and an infinitely distant directional light source with variable intensity. We place Gaussian priors over reflectances $\{r_i\}$, normals $\{\mathbf{n}_i\}$, and the light \mathbf{l} ; and set the parameters of the priors using training data. We additionally place a soft symmetry prior on the $\{r_i\}$ (the reflectance value on one side of the face should be close to its value on the other side) and on the $\{\mathbf{n}_i\}$ (normal vectors on each side should be approximately symmetric), reflecting our prior knowledge about faces. These symmetry priors can be added to the model in just a few lines of code, illustrating the way in which model-based methods lend themselves to rapid prototyping and experimentation.

Although this model is only a crude approximation to the true image formation process (e.g. it does not account for shadows or specularities), similar approximations have been found to be useful in prior work (Biswas et al., 2009; Biswas and Chellappa, 2010; Kemelmacher-Shlizerman and Basri, 2011). Additionally, if we can successfully develop algorithms that perform accurate and reliable inference in this class of models, we would then be able to increase its usefulness simply by updating it to reflect the true

image formation process more accurately. Note that even for a relatively small image of size 96×84 , the model contains over 48,000 latent variables and 56,000 factors, and as we will show below, standard message passing in the model routinely fails to converge to accurate solutions.

Consensus message passing. We use predictors at two levels in the model (see Fig. 6b) to tackle this problem. The first sends consensus messages to *each* reflectance pixel r_i , making it an instance of type B of CMP as described in Fig. 1b. Here, each consensus message is predicted using information from all the contextual messages from the z_i . We denote each of these predictors by Δ_i^r . The second predictor sends a consensus message to \mathbf{l} using information from all the messages from the s_i and is denoted by Δ^l . The first level of predictors effectively make a guess of the reflectance image from the denoised observation, and the second layer predictor produces an estimate of the light from the shading image (which is likely to be easier to do than directly from the observation). The reflectance predictors $\{\Delta_i^r\}$ are all powered by a single random forest, however the pixel position i is used as a feature that it can exploit to create location specific behaviour. The tree parameterization of the contextual messages \mathbf{c} for use in the reflectance predictor Δ_i^r also includes 16 features such as mean, median, max, min and gradients of a 21×21 patch around the pixel. The tree parameterization of the contextual messages for use in the lighting predictor Δ^l consists of means of the mean of the shading messages in 12×12 blocks. We deliberately use simple features to maintain generality but one could imagine the use of more specialized regressors for maximal performance.

Datasets. We experiment with the ‘Yale B’ and ‘Extended Yale B’ datasets (Georgiades et al., 2001; Lee et al., 2005). Together, they contain images of 38 subjects each with 64 illumination directions. We remove images taken with extreme light angles (azimuth or elevation ≥ 85 degrees) that are almost entirely in shadow, leaving around 45 images for each subject. Images are downsampled to 96×84 . There are no groundtruth normals or reflectances for this dataset, however it is common practice to create proxy groundtruths using photometric stereo, which we do using the code of Quéau et al. (2013). We use images from 22 subjects for training and test on the remaining 16 subjects.

Results. We begin by qualitatively assessing the different inference schemes. In Fig. 11 we show inference results for reflectance maps, normal maps and lights that are obtained following 100 iterations of message passing (VMP). For reflectance (Fig. 11b), we would like inference to produce estimates that match closely the groundtruth produced by photometric stereo (GT). We also display the reflectance estimates produced by the strong baseline of Biswas et al. (2009) for reference. We note that the baseline achieves excellent accuracy in regions with strong lighting, however it produces blurry estimates in regions under shadow.

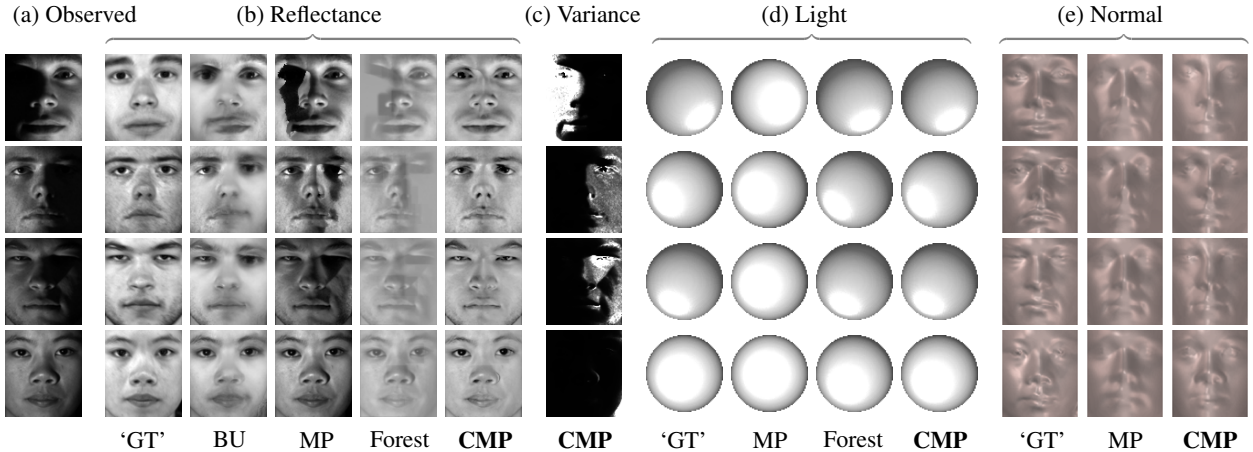


Figure 7: **A visual comparison of inference results.** For 4 randomly chosen test images, we show inference results obtained by competing methods. (a) Observed images. (b) Inferred reflectance maps. *GT* is the stereo estimate which we use as a proxy for groundtruth, *BU* is the bottom-up reflectance estimate of Biswas *et al.* (2009) and *Forest* is the consensus prediction. (c) The variance of the inferred reflectance estimate produced by CMP (normalized across rows). High variance regions correlate strongly with cast shadows. (d) Visualization of inferred light. (e) Inferred normal maps.

As can be seen in Fig. 11b (MP), standard variational message passing finds solutions that are highly inaccurate with continued presence of illumination and artefacts in areas of cast shadow. In contrast, inference using CMP produces artefact-free results that much more closely resemble the stereo groundtruths. Arguably CMP also improves over the baseline (Biswas *et al.*, 2009), since its estimates are not blurry in regions with cast shadows. This is made possible by the presence of symmetry priors in the model. Additionally, we note that the variance of the CMP inference for reflectance (Fig. 11c) correlates strongly with cast shadows in the observed images (*i.e.* the model is uncertain where it should be) suggesting that in future work it would be fruitful to have the notion of cast shadows explicitly built into the model. Figs. 11d and 11e show analogous results for lighting and normal maps, and Fig. 12 demonstrates CMP’s ability to robustly infer reflectance maps for images of a single subject taken under varying lighting conditions.

We use the task of subject recognition (using estimated reflectance) as a quantitative measure of inference accuracy, as it can be difficult to measure in more direct ways (*e.g.* RMSE strongly favours blurry predictions). The reflectance estimate produced by each algorithm is compared to all training subjects’ groundtruth reflectances and is assigned the label of its closest match. We have found this evaluation to reflect the quality of inference and we choose to use it for its simplicity. Fig. 13 shows the result of this experiment, both for real images and also synthetic images that were produced by taking the stereo groundtruths and adding artificial lighting (but with no cast shadows). We show analogous results for light in Fig. 14, where error is defined to be the cosine angle distance between the estimated light and the photometric stereo reference. First,

we note that standard variational message passing (MP) performs poorly, producing reflectance estimates that are much less useful for recognition than those from Biswas *et al.* (2009). Second, we note that CMP in the same model (both 1 stage and 2 stage versions) produces inferences that are significantly more useful downstream. The horizontal line labelled ‘Forest’ represents the accuracy of the consensus messages without any message passing, showing that the model-based fine-tuning provides a significant benefit. Finally, we highlight the fact that initializing light directly from the image and running message passing (Fig. 13, Init+MP) leads to worse estimates than CMP demonstrating the use of layered predictions as opposed to direct predictions from the observations. These results demonstrate that CMP helps message passing find better fixed points even in the presence of model mis-match (shadows) and make use of the full potential of the generative model.

5 Related Work

Inspiration for CMP stems from the kinds of distinctions that have been made for decades between so-called ‘intuitive’, bottom-up, fast inference techniques, and iterative ‘rational’ inference techniques (Hinton, 1990). CMP can be seen as an implementation of such ideas in the context of message passing, where the consensus messages form the ‘intuitive’ part of inference and the following standard message passing forms the ‘rational’ part. Analogues to intuitive and rational inference also exist for sampling, where bottom-up techniques are used to compute proposals for MCMC, leading to significant speedup in inference (Tu *et al.*, 2001; Stuhlmüller *et al.*, 2013; Jampani *et al.*, 2014). Rezende *et al.* (2014) and Kingma and Welling (2013) proposed techniques for learning the parameters of both the

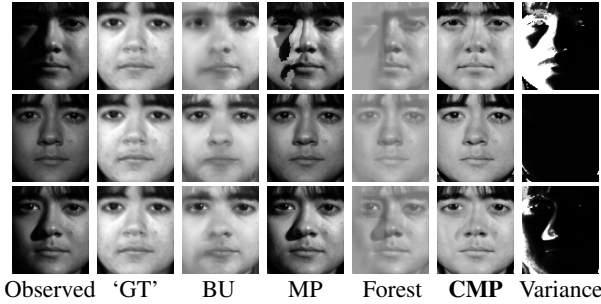


Figure 8: **Robustness to varying illumination.** Left to right: observed image, photometric stereo estimate (proxy for groundtruth), Biswas et al. (2009) estimate, VMP result, consensus forest estimate, CMP mean, and CMP variance.

generative model and the corresponding recognition model.

The idea of ‘learning to infer’ also has a long history. Early examples include Hinton et al. (1995), where a dedicated set of ‘recognition’ parameters are learned to drive inference. In more modern instances of such ideas (Munoz et al., 2010; Ross et al., 2011; Domke, 2011; Shapovalov et al., 2013; Munoz, 2013), message passing is performed by a sequence of predictions defined by a graphical model, and the predictors are jointly trained to ensure that the system produces correct labellings. However in these techniques the resulting inference procedure no longer corresponds to the original (or perhaps to any) graphical model. An important distinction of CMP is that the predictors fit completely within the framework of message passing and final inference results correspond to valid fixed points in the original model of interest.

Finally, we note recent works of Heess et al. (2013) and Es-lami et al. (2014) that make use of regressors (neural networks and random forests, respectively) to learn to pass EP messages. These works are concerned with reducing the computational cost of computing individual messages and do not make any attempt to change the accuracy or rate of convergence in message passing inference as a whole. In contrast, CMP learns to pass messages specifically with the aim of reducing the total number of iterations required for accurate inference in a given generative model.

6 Discussion

We have presented Consensus Message Passing and shown that it is a computationally efficient technique that can be used to improve the accuracy of message passing inference in a variety of vision models. The crux of the approach is to recognize the importance of global variables, and to take advantage of layered model structures commonly seen in vision to make rough estimates of their values.

The success of CMP depends on the accuracy of the random forest predictors. The design of forest features is not

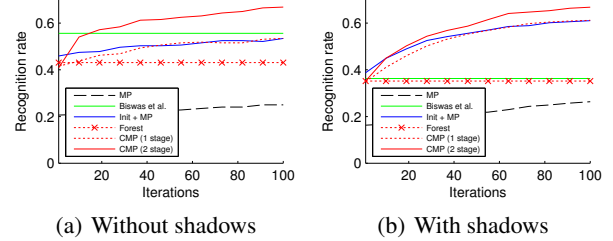


Figure 9: **Reflectance inference accuracy demonstrated through recognition accuracy.** CMP allows us to make use of the full potential of the generative model, thereby outperforming the competitive bottom-up method of Biswas et al. (2009).

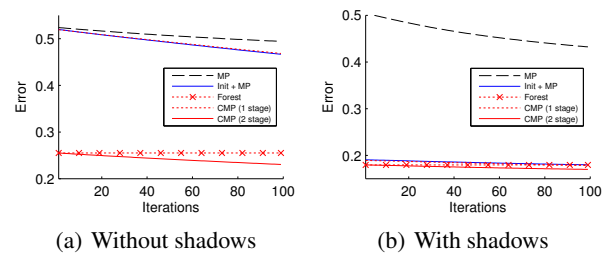


Figure 10: **Light inference accuracy.** The presence of cast shadows makes the direct prediction task easier, however CMP is accurate even in their absence.

yet completely automated, but we took care in this work to use generic features that can be applied to a broad class of problems. Our forests are implemented in an extensible manner, and we envisage building a library of them that one can choose from, simply by inspecting the data types of the contextual and target variables.

In future work, we would like to exploit the benefits of the CMP framework by applying it to more challenging problems from computer vision. Each of the examples in Sec. 4 can be extended in various ways, *e.g.* by making considerations for multiple objects, incorporating occlusion in the squares example and cast shadows in the faces example, or by developing more realistic priors. We are also seeking to understand in what other domains the application of our ideas may be fruitful.

More broadly, a major challenge in machine learning is that of enriching models in a scalable way. We continually seek to ask our models to provide interpretations of increasingly complicated, heterogeneous data sources. Graphical models provide an appealing framework to manage this complexity, but the difficulty of inference has long been a barrier to achieving these goals. The CMP framework takes us one step in the direction of overcoming this barrier.

Acknowledgements. We thank the anonymous reviewers, Tom Minka, Christopher Williams, Peter Gehler, Sebastian Nowozin and Andrew Fitzgibbon for their feedback and suggestions.

References

- Biswas, S., Aggarwal, G., and Chellappa, R. (2009). Robust estimation of albedo for illumination-invariant matching and shape recovery. *Pattern Analysis and Machine Intelligence*, 31(5):884–899.
- Biswas, S. and Chellappa, R. (2010). Pose-robust albedo estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2683–2690.
- Criminisi, A. and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company, Incorporated.
- Domke, J. (2011). Parameter learning with truncated message-passing. *Computer Vision and Pattern Recognition (CVPR)*, pages 2937–2943.
- Eslami, S. M. A., Tarlow, D., Kohli, P., and Winn, J. (2014). Just-In-Time Learning for Fast and Flexible Inference. In *Neural Information Processing Systems (NIPS)* 27, pages 154–162.
- Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence*, 23(6).
- Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. (2013). Annealing between distributions by averaging moments. In *Neural Information Processing Systems (NIPS)* 26, pages 2769–2777.
- Heess, N., Tarlow, D., and Winn, J. (2013). Learning to Pass Expectation Propagation Messages. In *Neural Information Processing Systems (NIPS)* 26, pages 3219–3227.
- Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46(1-2):47–75.
- Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161.
- Jampani, V., Nowozin, S., Loper, M., and Gehler, P. V. (2014). The Informed Sampler: A Discriminative Approach to Bayesian Inference in Generative Computer Vision Models. *arXiv preprint arXiv:1402.0859*.
- Kemelmacher-Shlizerman, I. and Basri, R. (2011). 3d face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence*, 33(2):394–405.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
- Lee, K.-C., Ho, J., and Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence*, 27(5):684–698.
- Minka, T. (2001). *Expectation Propagation for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology.
- Minka, T., Winn, J., Guiver, J., and Knowles, D. (2012). Infer.NET 2.5. Microsoft Research Cambridge. Website URL: <http://research.microsoft.com/infernet>.
- Munoz, D. (2013). *Inference Machines: Parsing Scenes via Iterated Predictions*. PhD thesis, The Robotics Institute, Carnegie Mellon University.
- Munoz, D., Bagnell, J. A., and Hebert, M. (2010). Stacked Hierarchical Labeling. In *European Conference on Computer Vision (ECCV)*, pages 57–70.
- Quéau, Y., Lauze, F., and Durou, J.-D. (2013). *Solving the uncalibrated photometric stereo problem using total variation*. Springer.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pages 1278–1286.
- Ross, S., Munoz, D., Hebert, M., and Bagnell, J. A. (2011). Learning Message-Passing Inference Machines for Structured Prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2737–2744.
- Shapovalov, R., Vetrov, D., and Kohli, P. (2013). Spatial Inference Machines. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2985–2992.
- Stuhlmüller, A., Taylor, J., and Goodman, N. D. (2013). Learning Stochastic Inverses. In *Neural Information Processing Systems (NIPS)* 27, pages 3048–3056.
- Tang, Y., Salakhutdinov, R., and Hinton, G. (2012). Deep lambertian networks. *arXiv preprint arXiv:1206.6445*.
- Teets, D. and Whitehead, K. (1999). The Discovery of Ceres: How Gauss Became Famous. *Mathematics Magazine*, 72(2):83–93.
- Tu, Z., Zhu, S.-C., and Shum, H.-Y. (2001). Image segmentation by data driven Markov chain Monte Carlo. In *International Conference on Computer Vision (ICCV)*, pages 131–138.
- Wang, Y., Zhang, L., Liu, Z., Hua, G., Wen, Z., Zhang, Z., and Samaras, D. (2009). Face relighting from a single image under arbitrary unknown lighting conditions. *Pattern Analysis and Machine Intelligence*, 31(11):1968–1984.
- Winn, J. and Bishop, C. M. (2005). Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694.
- Zhang, L. and Samaras, D. (2006). Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *Pattern Analysis and Machine Intelligence*, 28(3):351–363.

Consensus Message Passing for Layered Graphical Models

Supplementary Material

Varun Jampani[†] S. M. Ali Eslami[†], Daniel Tarlow, Pushmeet Kohli and John Winn
 MPI for Intelligent Systems, Tübingen Microsoft Research, Cambridge

1 Random regression forests for CMP

We wish to learn a mapping f from contextual messages \mathbf{c} to the consensus message m from training data $\{(\mathbf{c}_d, m_d)\}_{d=1\dots D}$. This is challenging since the inputs and outputs of the regression problem are both messages (*i.e.* distributions), and special care needs to be taken to account for this fact. We follow closely the methodology of Eslami et al. (2014), who use random forests to predict outgoing messages from a factor given the incoming messages to it. For a review of forests see (Criminisi and Shotton, 2013).

In approximate message passing (*e.g.* EP; Minka, 2001 and VMP; Winn and Bishop, 2005), messages can be represented using only a few numbers, *e.g.* a Gaussian message can be represented by its natural parameters. We represent the contextual messages \mathbf{c} collectively, in two different ways: the first is a concatenation of the parameters of its constituent messages which we call the ‘regression parameterization’ and denote by \mathbf{r}_c ; and the second is a vector of features computed on the set which we call the ‘tree parameterization’ and denote by \mathbf{t}_c . This parametrization typically contains features of the set as a whole (*e.g.* moments of their means). We represent the outgoing message m by a vector of real valued numbers \mathbf{r}_m .

Prediction model. Each leaf node is associated with a subset of the labelled training data. During testing, a previously unseen set of contextual messages represented by \mathbf{t}_c traverses the tree until it reaches a leaf which by construction is likely to contain similar training examples. We therefore use the statistics of the data gathered in that leaf to predict the consensus message with a multivariate regression model of the form: $\mathbf{r}_m = \mathbf{W} \cdot \mathbf{r}_c + \epsilon$ where ϵ is a vector of normal error terms. We use the learned matrix of coefficients \mathbf{W} at test time to make predictions $\bar{\mathbf{r}}_m$ for each \mathbf{r}_c . To recap, \mathbf{t}_c is used to traverse the contextual messages down to leaves, and \mathbf{r}_c is used by a linear regressor to predict the parameters \mathbf{r}_m of the consensus message.

Training objective function. The optimization of the split functions proceeds in a greedy manner. At each node j , depending on the subset of the incoming training set \mathcal{S}_j we learn the function that ‘best’ splits \mathcal{S}_j into the training sets corresponding to each child, \mathcal{S}_j^L and \mathcal{S}_j^R , *i.e.* the parameters of the split criterion $\tau_j = \operatorname{argmax}_{\tau \in \mathcal{T}_j} I(\mathcal{S}_j, \tau)$. This optimization is performed as a search over a discrete set \mathcal{T}_j of a random sample of possible parameter settings. The objective function I is:

$$I(\mathcal{S}_j, \tau) = -E(\mathcal{S}_j^L, \mathbf{W}^L) - E(\mathcal{S}_j^R, \mathbf{W}^R), \quad (1)$$

where \mathbf{W}^L and \mathbf{W}^R are the parameters of the regression models corresponding to the left and right training sets \mathcal{S}_j^L and \mathcal{S}_j^R , and E is the ‘fit residual’ as defined in (Eslami et al., 2014). In simple terms, this objective function splits the training data at each node in a way that the relationship between the incoming and outgoing messages is well captured by the regression in each child.

Ensemble model. During testing, a set of contextual messages simultaneously traverses every tree in the forest from their roots until it reaches their leaves. Combining the predictions into a single forest prediction might be done by averaging the parameters $\bar{\mathbf{r}}_m^t$ of the predicted messages \bar{m}^t by each tree t , however this would be sensitive to the chosen parameterization. Instead we compute the moment average \bar{m} of the distributions $\{\bar{m}^t\}$ by averaging the first few moments of the predictions across trees, and solving for the distribution parameters which match the averaged moments (see *e.g.* Grosse et al., 2013).

[†]The first two authors contribute equally to this work.

2 Results on the face problem

2.1 Qualitative results

Figure 11 shows inference results for reflectance maps, normal maps and lights for randomly chosen test images, and Figure 12 shows reflectance estimation results on multiple images of the same subject produced under different illumination conditions. Consensus message passing is able to produce reflectance estimates that are closer to the photometric stereo groundtruth across subjects and across different illumination conditions.

2.2 Quantitative results

Figure 13 shows quantitative results for both real images from ‘Yale B’ and ‘Extended Yale B’ datasets (Georghiades et al., 2001; Lee et al., 2005) and synthetic shadowless images. The synthetic shadowless images were created using the same light, reflectance and normal map statistics as that of images in the real dataset (however estimated using photometric stereo (Quéau et al., 2013)). Subject recognition results indicate superior performance of CMP in comparison to other baselines in both real and synthetic image settings.

Figure 14 shows the quantitative results of light inference using the different inference techniques. We use the cosine angle distance between the estimated light and the photometric stereo groundtruth ($\text{error} = \cos^{-1}(\hat{\mathbf{l}}_{\text{est}} \cdot \hat{\mathbf{l}}_{\text{ps}})$) as an error metric. Here, $\hat{\mathbf{l}}_{\text{est}}$ is a unit vector in the same direction as the mean of the posterior light estimate of CMP and $\hat{\mathbf{l}}_{\text{ps}}$ is a unit vector in the same direction as the corresponding photometric stereo groundtruth. Again, these results indicate the superior performance of CMP in comparison to other baselines in both real and synthetic image settings.

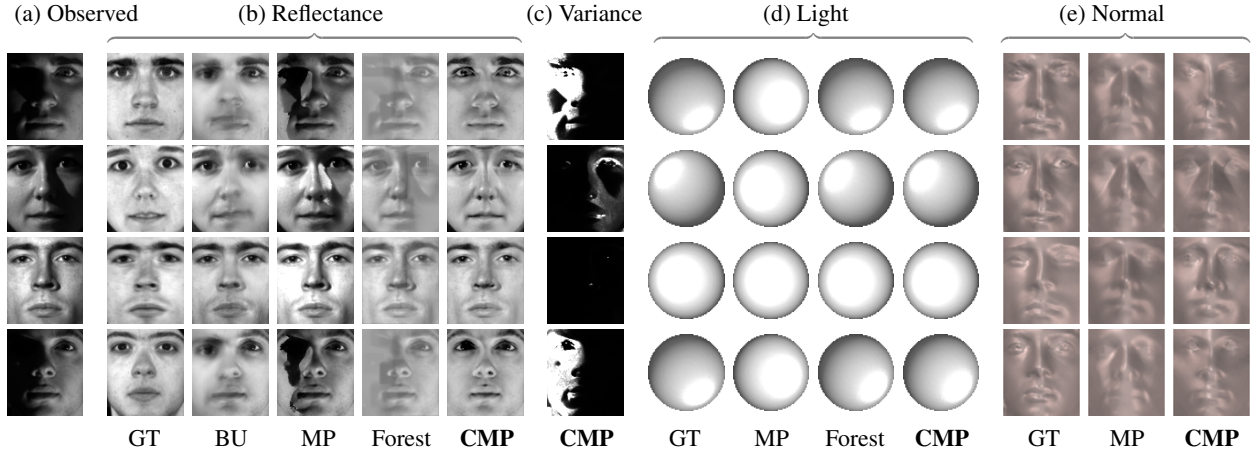


Figure 11: **A visual comparison of inference results.** For 4 randomly chosen test images, we show inference results obtained by competing methods. (a) Observed images. (b) Inferred reflectance maps. *GT* is the photometric stereo groundtruth, *BU* is the Biswas *et al.* (2009) reflectance estimate and *Forest* is the consensus prediction. (c) The variance of the inferred reflectance estimate produced by CMP (normalized across rows). High variance regions correlate strongly with cast shadows. (d) Visualization of inferred light directions. (e) Inferred normal maps.

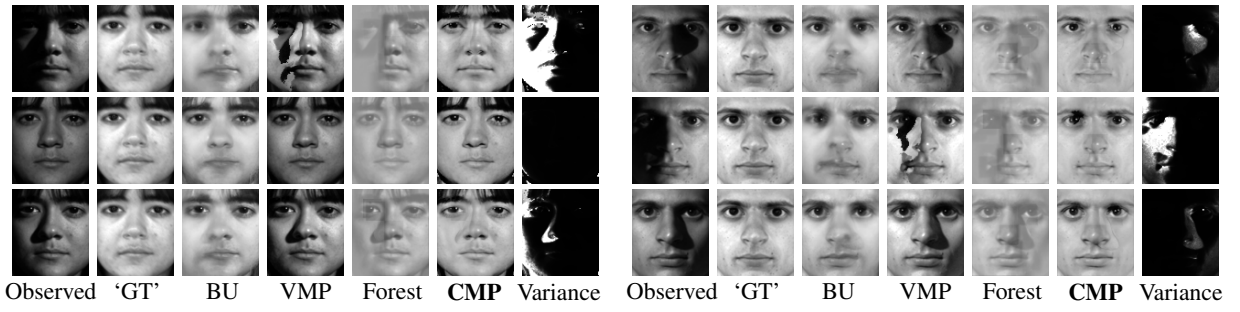


Figure 12: **Robustness to varying illumination.** Reflectance estimation on two subject images with varying illumination. Left to right: observed image, photometric stereo estimate which is used as a proxy for groundtruth, bottom-up estimate of Biswas et al. (2009), VMP result, consensus forest estimate, CMP mean, and CMP variance.

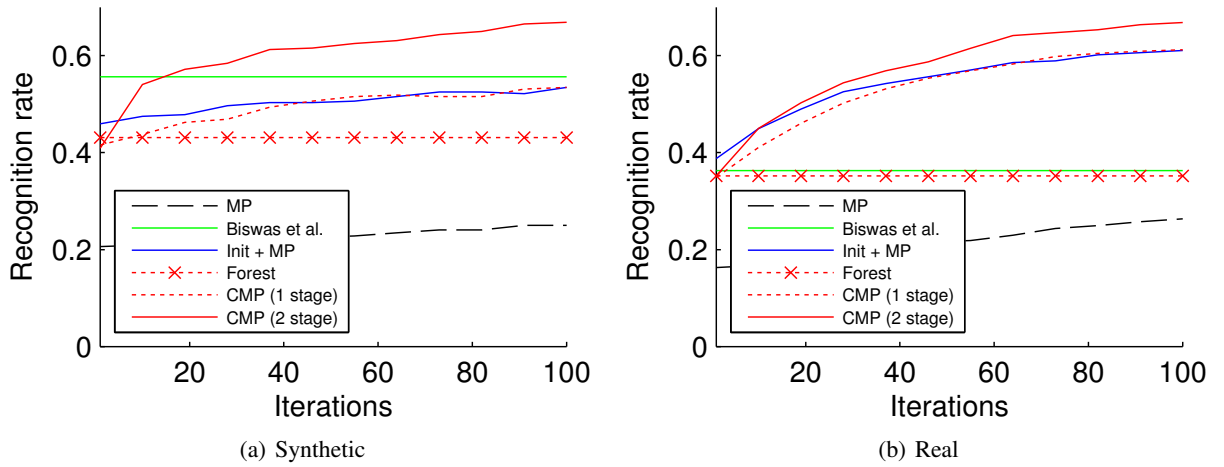


Figure 13: **Reflectance inference accuracy.** Results have been averaged over all images of test subjects. (a) Synthetic, shadowless images. (b) Real images.

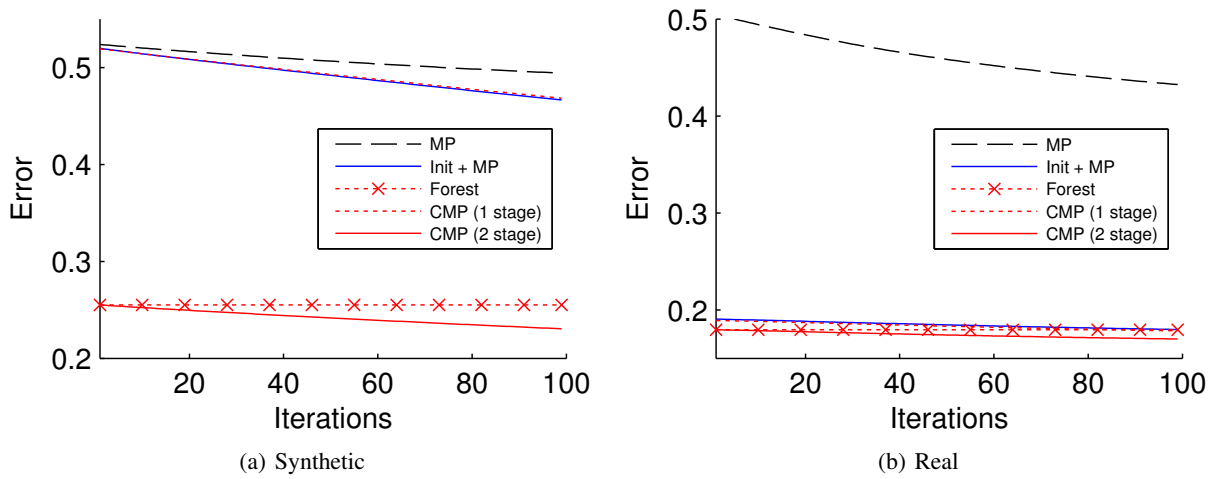


Figure 14: **Light inference accuracy.** Results have been averaged over all images of test subjects. (a) Synthetic, shadowless images. (b) Real images.